

Robust Nonnegative Sparse Recovery and the Nullspace Property of 0/1 Measurements

Richard Küng*, Peter Jung†

*Institute for Theoretical Physics, University of Cologne

†Communications and Information Theory Group, Technische Universität Berlin

rkueng@thp.uni-koeln.de, peter.jung@tu-berlin.de

Abstract

We investigate recovery of nonnegative vectors from non-adaptive compressive measurements in the presence of noise of unknown power. It is known in literature that under additional assumptions on the measurement design recovery is possible in the noiseless setting with nonnegative least squares without any regularization. We show that such known uniqueness results carry over to the noisy setting. We present guarantees which hold instantaneously by establishing the relation to the robust nullspace property. As an important example, we establish that an $m \times n$ random iid. 0/1-valued Bernoulli matrix has with overwhelming probability the robust nullspace property for $m = \mathcal{O}(s \log(n))$ and is applicable in the nonnegative case. Our analysis is motivated by applications in wireless network activity detection.

I. INTRODUCTION

Recovery of lower complexity objects by observations far below the Nyquist rate has applications in physics, applied math, and many engineering disciplines. Moreover, it is one of the key tools for facing challenges in data processing (like big data and the Internet of Things), wireless communications (the 5th generation of the mobile cellular network) and large scale network control. Compressed Sensing (CS), with its origin in the recovery of sparse or compressible vectors has, in particular, stimulated the research community to investigate further directions of compressibility and low-dimensional structures which allow the recovery from low-rate samples and with efficient algorithms. In many applications, the objects of interest exhibit further structural constraints which should be exploited in reconstruction algorithms. Take, for instance, the following setting which appears naturally in communication protocols: the components of sparse information carrying vectors are taken from a finite alphabet or the data vectors are lying in specific subspaces. Similarly, in network traffic estimation and anomaly detection from end-to-end measurements, the parameters are restricted to particular lower-dimensional domains. Finally, the signals occurring in imaging problems are typically constrained to non-negative intensities.

Our work is partially inspired by the task of identifying sparse network activation patterns in a large-scale asynchronous wireless network: suppose that, in order to indicate its presence, each active device node transmits an individual sequence into a noisy wireless channel. All such sequences are multiplied with individual, but unknown,

channel amplitudes¹ and finally superimposed at the receiver. The receiver's task then is to detect all active devices and the corresponding channel amplitudes from this global superposition (note that each device is uniquely characterized by the sequence it transmits). This problem can be re-cast as the task of estimating non-negative sparse vectors from noisy linear observations.

Such non-negative and sparse structures also arise naturally in certain empirical inference problems, like network tomography [1], [2], statistical tracking (see e.g. [3]) and compressed imaging of intensity patterns [4]. The underlying mathematical problem has received considerable attention in its own right [5], [6], [7], [8], [9]. It has been shown that measurement matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ coming from *outwardly s -neighborly polytopes* [10] and matrices $\mathbf{A} \in \mathcal{M}^+$ whose *row span intersects the positive orthant*² [11] maintain an intrinsic uniqueness property for non-negative, s -sparse vectors even in the underdetermined setting ($m < n$). Such uniqueness properties in turn allow for entirely avoiding CS algorithms in the reconstruction step. From an algorithmic point of view, this is highly beneficial. However, all the statements mentioned above are manifestly focussed on idealized scenarios, where no noise is present in the sampling procedure.

Motivated by device detection, we shall overcome this idealization and devise recovery protocols that are robust towards any form of additive noise. Our results have the added benefit that no a-priori bound on the noise step is required in the reconstruction algorithm.

A. Main Results

Let us introduce some notation and then state our main findings. Throughout our work we endow \mathbb{R}^n with the partial ordering induced by the nonnegative orthant, i.e. $\mathbf{x} \leq \mathbf{z}$ if and only if $x_i \leq z_i$ for all $1 \leq i \leq n$. Here, $x_i = \langle \mathbf{e}_i, \mathbf{x} \rangle$ are the components of \mathbf{x} with respect to the standard basis $\{\mathbf{e}_i\}_{i=1}^n$. Similarly, we write $\mathbf{x} < \mathbf{z}$ if strict inequality holds in each component. Consequently, we write $\mathbf{x} \geq \mathbf{0}$ to indicate that \mathbf{x} is (entry-wise) nonnegative. For $1 \leq p \leq \infty$, we denote the ℓ_p -norms of vectors by $\|\cdot\|_{\ell_p}$ and $\|\cdot\|$ is the usual operator/matrix norm. The sparsity of a vector \mathbf{x} is denoted by $\|\mathbf{x}\|_{\ell_0} := |\text{supp}(\mathbf{x})| \leq s$ where $\text{supp}(\mathbf{x}) := \{i : x_i \neq 0\}$ is its support in the standard basis.

Mathematically, we are interested in recovering sparse, nonnegative vectors $\mathbf{x} \in \mathbb{R}^n$ from $m \ll n$ erroneous linear measurements of the form $y_i = \mathbf{a}_i^T \mathbf{x} + e_i$. Here, the vectors $\mathbf{a}_i \in \mathbb{R}^n$ model the different measurement operations and e_i is additive noise of arbitrary size and nature. By encompassing all \mathbf{a}_i 's as rows of a sampling matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and defining $\mathbf{y} = (y_1, \dots, y_m)^T$, as well as $\mathbf{e} = (e_1, \dots, e_m)^T$, such a sampling procedure can succinctly be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \quad (1)$$

Several conditions on \mathbf{A} are known which are sufficient to ensure that a sparse vector \mathbf{x} can be robustly estimated from measurements \mathbf{y} . A famous condition is the *restricted isometry property* (RIP). A matrix $\tilde{\mathbf{A}}$ is said to

¹This can be justified under certain assumptions like pre-multiplications using channel reciprocity in time-division multiplexing.

²See (7) below for a precise definition.

be s -RIP, if it acts almost isometrically on s -sparse vectors, meaning that there exists a $\delta_s \in [0, 1)$ such that $|\|\tilde{\mathbf{A}}\mathbf{x}\|_{\ell_2}^2 - \|\mathbf{x}\|_{\ell_2}^2| \leq \delta_s \|\mathbf{x}\|_{\ell_2}^2$ for all s -sparse \mathbf{x} . When dealing with random matrices \mathbf{A} , one has also to distinguish between uniform and non-uniform guarantees³. It is a well-known fact that RIP is only sufficient but not necessary for uniform recovery. Overcoming this asymmetry, the notion of a *nullspace property* assures that no s -sparse vectors lie in the kernel of \mathbf{A} . Hence, the NSP is both a sufficient and necessary condition for recovery. Proving that (1) indeed allows for *robustly* recovering any s -sparse \mathbf{x} in the presence of noise therefore is equivalent to establishing that \mathbf{A} obeys a *robust nullspace property of order s* (NSP) [12, Chapter 4]. Our first main technical contribution is a substantial strengthening of the implications of such an NSP for reconstructing nonnegative sparse vectors:

Theorem 1. *Suppose that \mathbf{A} obeys the NSP of order $s \leq n$ from Def. 3 and moreover admits a strictly-positive linear combination of its rows ($\mathbf{A} \in \mathcal{M}^+$, i.e., $\exists \mathbf{t} \in \mathbb{R}^m$ such that $\mathbf{w} = \mathbf{A}^T \mathbf{t} > \mathbf{0}$). Then, the following bound holds for any s -sparse $\mathbf{x} \geq \mathbf{0}$ and any $\mathbf{z} \geq \mathbf{0}$:*

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{D'}{\sqrt{m}} \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|_{\ell_2}. \quad (2)$$

The constant D' only⁴ depends on the quality of NSP and the conditioning of the strictly positive vector \mathbf{w} .

We are interested in retrieving \mathbf{x} from the measurements \mathbf{y} in (1). Inserting this equation into the r.h.s of (2) and applying the triangle inequality reveals

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{D'}{\sqrt{m}} (\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) \quad \forall \mathbf{z} \geq \mathbf{0}.$$

This data-dependent bound suggests to minimize its right hand side over the “free parameter” $\mathbf{z} \geq \mathbf{0}$ in order to get an estimator $\mathbf{x}^\#$ of \mathbf{x} , i.e.

$$\mathbf{x}^\# = \arg \min_{\mathbf{0} \leq \mathbf{z} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2}. \quad (3)$$

This is a simple *nonnegative least squares regression* (NNLS) that does not require any assumptions on the noise \mathbf{e} . Since the target vector \mathbf{x} is itself nonnegative and therefore a feasible point of (3), we can furthermore conclude

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^\#\|_{\ell_2} &\leq \frac{D'}{\sqrt{m}} (\arg \min_{\mathbf{z} \geq \mathbf{0}} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) \\ &\leq \frac{D'}{\sqrt{m}} (\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) = \frac{2D'}{\sqrt{m}} \|\mathbf{e}\|_{\ell_2}, \end{aligned} \quad (4)$$

where we have once more resorted to (1). Consequently, Theorem 1 assures that solving (3) yields an estimator of any s -sparse vector $\mathbf{x} \geq \mathbf{0}$. Moreover, this estimator is robust towards additive noise in the sampling process. Such a recovery guarantee is (up to multiplicative constants) as strong as existing ones for different reconstruction algorithms, including the LASSO and Dantzig selectors, as well as basis pursuit denoising (BPDN) (see [12] and references therein). However, on the contrary to them, algorithms for solving (3) require neither an explicit a-priori bound

³Non-uniform guarantees hold w.h.p. for priorly fixed vectors \mathbf{x} , while uniform guarantees assure recovery of all s -sparse vectors simultaneously. RIP is an example for the latter.

⁴See Theorem 4 below for explicit dependencies.

$\eta \geq \|\mathbf{e}\|_{\ell_2}$ on the noise, nor an $\|\cdot\|_{\ell_1}$ regression term. This *remarkable simplicity* is caused by the non-negativity constraint $\mathbf{z} \geq \mathbf{0}$ and the geometric restrictions it imposes. Also, these assertions stably remain true, if we consider approximately sparse target vectors instead of perfectly sparse ones (see Theorem 4 below).

In order to underline the applicability of Theorem 1, we consider nonnegative 0/1-Bernoulli sampling matrices and prove that they meet the requirements of said statement with high probability (w.h.p.).

Theorem 2. *Let \mathbf{A} be a sampling matrix whose entries are independently chosen from a 0/1-Bernoulli distribution with parameter $p \in [0, 1]$, i.e. $\Pr[1] = p$ and $\Pr[0] = 1 - p$. Fixing $s \leq n$ and setting*

$$m \geq \frac{C}{(p(1-p))^2} s \left(\log(n) + \frac{p}{1-p} \right) \quad (5)$$

assures that \mathbf{A} obeys the NSP from Definition 3 and the vector $\mathbf{w} := \mathbf{A}^T \left(\frac{1}{pm} \mathbf{1} \right)$ obeys $\max_{1 \leq i \leq n} |w_i - 1| < 1/2$ (and is thus strictly positive) with probability at least $1 - (n+1)e^{-C' p^2 (1-p)^2 m}$.

Combining this statement with (4) implies that w.h.p. such Bernoulli matrices allow for uniformly and stably reconstructing sparse, nonnegative vectors \mathbf{x} via Alg. (3). We demonstrate this numerically in Figure 1. Up to our knowledge, this is the first rigorous proof that 0/1-matrices tend to obey a strong version of the nullspace property. The challenging difference to existing NSP and RIP results is the fact that the individual random entries of \mathbf{A} are not centered, ($\mathbb{E}[\mathbf{A}_{k,j}] = p \neq 0$). Thus, the covariance matrix of \mathbf{A} admits a condition number of $\kappa(\mathbb{E}[\mathbf{A}^T \mathbf{A}]) = 1 + \frac{pn}{1-p}$, which underlines the ensemble's anisotropy. Traditional proof techniques, like establishing an RIP, are either not applicable in such a setting, or yield sub-optimal results [13], [14]. This is not true for Mendelson's small ball method [15], [16] (see also [17]), which we employ in our proof. This method is a strong general purpose tool whose applicability only requires row-wise independence, not centeredness. In the conceptually similar problem of reconstructing low rank matrices from rank-one projective measurements (which arises e.g. from the PhaseLift approach for phase retrieval [18]), applying this technique allowed for establishing strong null space properties, despite a similar degree of anisotropy in the sampling model [19]. A detailed survey of the applicability of Mendelson's small ball method for compressed sensing was recently presented in [20].

Organization of the Paper: In Section II we explain our motivating application in more detail and rephrase activity detection as a nonnegative sparse recovery problem. Then, we provide an overview on prior work and known results regarding this topic. In Section III we show that recovery guarantees in the presence of noise are governed here by the *robust nullspace property* (see here [12]) *under nonnegative constraints* which hasn't been fully analyzed so far in literature. It turns out that this property assures that any nonnegative s -sparse vector can be robustly recovered using conventional nonnegative least-squares. We stress out that such an algorithm requires *no a priori-knowledge* on the norm of the noise vector. Finally, in Section IV we analyze binary measurements matrices having iid. random 0/1-valued entries and we show that with overwhelming probability such matrices admit the robust nullspace property on nonnegative vectors. We obtain this result make use of a recent tool, known as "Mendelson's small ball method" which has already used by one of the authors in a related matrix recovery problem [19].

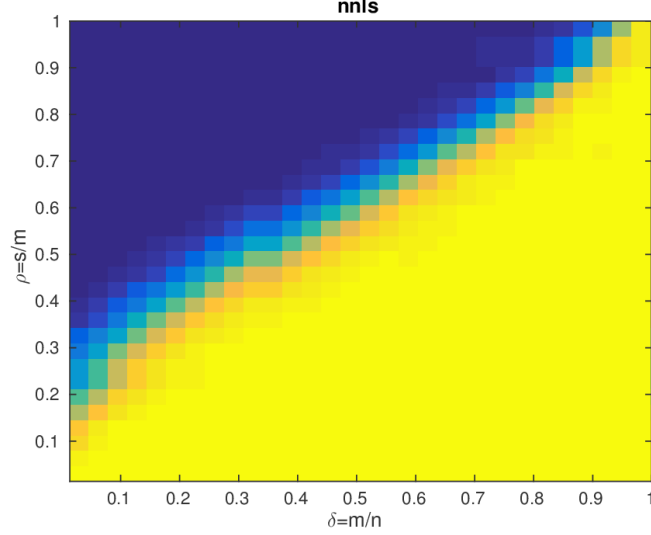


Fig. 1: Phase transition for NNLS in (3) – for iid. 0/1-Bernoulli measurement matrices in the noiseless case. More details are given in Section V.

II. SYSTEM MODEL AND PROBLEM STATEMENT

A. Activity Detection in Wireless Networks

Let $\mathbf{A} = (\mathbf{s}_j)_{j=1}^n \in \mathbb{R}^{m \times n}$ be a matrix with n real columns $\mathbf{s}_j \in \mathbb{R}^m$. In our network application [21], the columns \mathbf{s}_j are the individual sequences of length m transmitted by the active devices. These sequences are transmitted simultaneously and each of them is multiplied by an individual amplitude that depends on transmit power and other channel conditions. In practice this can be achieved for example using the channel reciprocity principle in time-division multiplexing so that the devices have knowledge about the complex channel coefficients and perform a corresponding pre-multiplication to correct for the phase. At a single receiver, all these modulated sequences are superimposed, because a single wireless medium is shared by all devices. We model such a situation by an unknown non-negative vector $\mathbf{0} \leq \mathbf{x} \in \mathbb{R}^n$, where $x_i > 0$ indicates that a device with sequence i is active with amplitude x_i ($x_i = 0$ implies that a device is inactive). We point out that, due to path loss in the channel, the individual received amplitudes x_i of each active device are unknown to the receiver as well. Here, we focus on networks that contain a large number n of registered devices, but, at any time, only a small unknown fraction, say $s \ll n$, of these devices are active.

Communicating activity patterns, that is $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$, and the corresponding list of received amplitudes/powers ($\mathbf{x} \geq \mathbf{0}$ itself) in a traditional way would require an $\mathcal{O}(n)$ resources to perform this task. We aim therefore for a reduction of the signaling time m by exploiting the facts that (i) $\mathbf{x} \geq \mathbf{0}$ is non-negative and (ii) the vector \mathbf{x} is s -sparse, i.e. $\|\mathbf{x}\|_{\ell_0} \leq s$. Hence, we assume that $s \leq m \ll n$. Obviously, in such a scenario the resulting system of linear equations cannot be directly inverted. A reasonable approach towards recovery is to consider the

program:

$$\arg \min \|\mathbf{x}\|_{\ell_0} \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{y} \ \& \ \mathbf{x} \geq \mathbf{0}$$

Combinatorial problems of this type are infamous for being NP-hard in general. A common approach to circumvent this obstacle is to consider convex relaxations. A prominent relaxation is to replace $\|\cdot\|_{\ell_0}$ with the ℓ_1 -norm. The resulting algorithm can then be re-cast as an efficiently solvable linear program. However, such approaches become more challenging when robustness towards additive noise is required, in particular if the type and the strength of the noise is itself unknown. In our application, noisy contributions inevitable arises due to quantization, thermal noise and other interferences. If the noisy measurements are of the form (1) (i.e. $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ where the vector \mathbf{e} is an additive distortion) a well-known modification is then to consider

$$\arg \min \|\mathbf{x}\|_{\ell_1} \quad \text{s.t.} \quad \|\mathbf{Ax} - \mathbf{y}\|_{\ell_2} \leq \eta \ \& \ \mathbf{x} \geq \mathbf{0}. \quad (6)$$

While this is not a linear problem anymore, it is still convex and is computationally tractable. In practice further modifications are necessary to solve such problems also sufficiently fast and efficiently, see [21]. However, having access to an apriori bound η on $\|\mathbf{e}\|_{\ell_2}$ is essential for (i) posing this problem and (ii) solving it using certain algorithms (stopping conditions etc.). Suppose, for instance, that \mathbf{e} is iid normal distributed. Then $\|\mathbf{e}\|_{\ell_2}^2$ admits a χ^2 -distribution of order m and feasibility is assured w.h.p., when taking η in terms of second moments. However, much less is known for different noise distributions or for situations, where second moment information about the noise is challenging to acquire.

One option to tackle problems of this kind is to establish a *quotient property* for the measurement matrix \mathbf{A} [12]. However, this property is geared towards Gaussian measurements and it is challenging to establish it, if \mathbf{A} follows a different random model. We shall show below that, interestingly, requiring $\mathbf{A} \in \mathcal{M}^+$ instead allows for drawing similar conclusions.

B. Prior Work on Recovery of Nonnegative Sparse Vectors

One of the first works in the noiseless setting is due to Donoho et al. [4] on the “nearly black object”. It furthers understanding of the “maximum entropy inversion” method to recover sparse (nearly-black) images in radio astronomy. In [10], Donoho and Tanner investigated this subject more directly. The question is, when \mathbf{A} intrinsically ensures that for each s -sparse $\mathbf{x}^{(0)}$ only one solution is feasible:

$$\{\mathbf{y} \mid \mathbf{Ax} = \mathbf{Ax}^{(0)} \ \& \ \mathbf{x} \geq \mathbf{0}\} = \{\mathbf{x}^{(0)}\}$$

At the center of their work is the notion of *outwardly s -neighborly polytopes*. Assume w.l.o.g. that all columns \mathbf{s}_j of \mathbf{A} are non-zero and define their convex hull

$$P_{\mathbf{A}} := \text{conv}(\mathbf{s}_1, \dots, \mathbf{s}_n).$$

This polytope is called *s -neighborly*, if every set of s vertices spans a face of $P_{\mathbf{A}}$. If this is the case, the polytope $P_{\mathbf{A}}^0 := \text{conv}(P_{\mathbf{A}} \cup \{\mathbf{0}\})$ is called then *outwardly s -neighborly*. They then move on to prove that the solution to

$$\arg \min \|\mathbf{x}\|_{\ell_0} \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{y}$$

is unique if and only if $P_{\mathbf{A}}^0$ is outwardly s -neighborly (see [10]). Another notion is the set of full-rank $m \times n$ -matrices having *intersection of its row space with the positive orthant* as introduced in [11]:

$$\mathcal{M}^+ = \{\mathbf{A} : \exists \mathbf{t} \in \mathbb{R}^m \mathbf{A}^* \mathbf{t} > 0\}. \quad (7)$$

Note that both structures are related in the sense that $\mathbf{A} \in \mathcal{M}^+$, if and only if $0 \notin P_{\mathbf{A}}$ [22]. In [11] Bruckstein et al. investigated the recovery of nonnegative vectors by (6) and modifications of OMP using a coherence-based approach. They obtained numerical evidence for unique recovery in the regime $s = \mathcal{O}(\sqrt{n})$. Later, Wang and coauthors [22] have analyzed non-negativity priors for vector and matrix recovery using an RIP-based analysis. Concretely, they translated the well-known RIP-result of random iid. ± 1 -Bernoulli matrices (see for example [23]) to 0/1-measurements in the following way. Let

$$\mathbf{1}_n := (1, \dots, 1)^T$$

denote the “all-ones” vector in \mathbb{R}^n . Perform measurements using an $(m+1) \times n$ matrix $\mathbf{A}^1 = (\mathbf{1}_n^T | \mathbf{A}^T)^T$ which consists of an all-ones row $\mathbf{1}_n$ appended by a random iid. 0/1-valued $m \times n$ matrix \mathbf{A} . By construction, the first noiseless measurement on a nonnegative vector \mathbf{x} returns its ℓ_1 -norm $\|\mathbf{x}\|_{\ell_1} = \langle \mathbf{1}_n, \mathbf{x} \rangle$. Rescaling and subtracting this value from the m remaining measurements then results in ± 1 -measurements. This insight allows for an indirect nullspace characterization of \mathbf{A} in terms of the RIP-constant δ_{2s} (see above, paragraph below (1)) of iid ± 1 -Bernoulli random matrices $\tilde{\mathbf{A}}$. More precisely [24]: For each $\mathbf{v} \in \mathcal{N}(\tilde{\mathbf{A}})$ in the nullspace $\mathcal{N}(\tilde{\mathbf{A}})$ of $\tilde{\mathbf{A}}$, an (ℓ_1, ℓ_1) -nullspace property is valid. Mathematically this means

$$\|\mathbf{v}_S\|_{\ell_1} \leq \frac{\sqrt{2}\delta_{2s}}{1 - \delta_{2s}} \|\mathbf{v}_{\bar{S}}\|_{\ell_1} \quad (8)$$

for all $\mathbf{v} \in \mathcal{N}(\tilde{\mathbf{A}})$ and $|S| \leq s$. Combining this with $\mathcal{N}(\mathbf{A}^1) \subset \mathcal{N}(\tilde{\mathbf{A}})$ then allows for proving unique recovery in regime $s = \mathcal{O}(n)$ with overwhelming probability.

However, so far, all these results manifestly focus on noiseless measurements. Thus, the robustness of these approaches towards noise corruption needs to be examined. Foucart, for instance, considered the ℓ_1 -squared nonnegative regularization [9]:

$$\min_{\mathbf{x} \geq 0} \|\mathbf{x}\|_{\ell_1}^2 + \lambda^2 \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_2}^2 \quad (9)$$

which can be re-cast as nonnegative least-squares problem. He then showed that for stochastic matrices⁵ the solution of (9) converges to the solution of (6) for $\lambda \rightarrow \infty$.

Contrary to this, we aim at establishing even stronger recovery guarantees that, among other things, *do not require an a priori noise bound*. We have already mentioned that the quotient property would assure such bounds for Gaussian matrices in the optimal regime. But $m \times n$ Gaussian matrices fail to be in \mathcal{M}^+ with probability approaching one as long as $\lim_{n \rightarrow \infty} m/n < \frac{1}{2}$ [22]. On the algorithmic side, there exists variations of certain regression methods where the regularization parameter can be chosen independent of the noise power – see the overview article [25]

⁵Recall that a matrix is stochastic, if all entries are non-negative and all columns sum up to one.

for more details. For the LASSO selector, in particular, such modifications are known as the “scaled LASSO” and “square root LASSO” [26], [27].

Non-negativity as a further structural constraint has also been investigated in the statistics community. But these works focus on the averaged case with respect to (sub-)Gaussian additive noise, whereby we consider instantaneous guarantees. Slawski and Hein [8], as well as Meinshausen [7] have recently investigated this averaged setting.

III. NULLSPACE PROPERTY WITH NONNEGATIVE CONSTRAINTS

We use the following notation. For a given vector $\mathbf{x} \in \mathbb{R}^n$ and a set $S \subset [n] := [1 \dots n]$ we denote by \mathbf{x}_S the vector containing only the coefficients of \mathbf{x} in S . Let \bar{S} the complement of S in $[1 \dots n]$ such that $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_{\bar{S}}$. The ℓ_q -error of the best s -term approximation of a vector \mathbf{x} will be denoted by $\sigma_k(\mathbf{x})_{\ell_q}$. The well-known convex relaxation of the ℓ_0 -minimization with respect to an apriori ℓ_2 -bound η on the residual $\mathbf{Ax} - \mathbf{y}$ is *basis pursuit denoising* (BPDN):

$$\Delta_\eta(\mathbf{y}) = \arg \min \|\mathbf{x}\|_{\ell_1} \quad \text{s.t.} \quad \|\mathbf{Ax} - \mathbf{y}\|_{\ell_2} \leq \eta \quad (10)$$

A. The robust nullspace property

Let us recall the definition of the ℓ_2 -robust nullspace property with respect to the ℓ_2 -norm [12, Def. 4.21].

Definition 3 (ℓ_2 -robust nullspace property). *A $m \times n$ matrix \mathbf{A} satisfies the ℓ_2 -robust null space property of order s with parameters $\rho \in (0, 1)$ and $\tau > 0$, if:*

$$\|\mathbf{v}_S\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_{\bar{S}}\|_{\ell_1} + \tau \|\mathbf{Av}\|_{\ell_2} \quad \text{for all } \mathbf{v} \in \mathbb{R}^n \quad (11)$$

holds for all $S \subset [n]$ with $|S| \leq s$.

The ℓ_2 -robust nullspace property order s (s -NSP) allows for drawing the following conclusion [12, Theorem 4.25]: for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{C}{\sqrt{s}} (\|\mathbf{z}\|_{\ell_1} - \|\mathbf{x}\|_{\ell_1} + 2\sigma_s(\mathbf{x})_{\ell_1}) + D \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \quad (12)$$

is true, where $C = \frac{(1+\rho)^2}{1-\rho}$ and $D = \frac{3+\rho}{1-\rho}\tau$. Replacing \mathbf{z} with the BPDN minimizer $\mathbf{x}_\eta = \Delta_\eta(\mathbf{y})$ from (10) for the sampling model $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ then implies

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_\eta\|_{\ell_2} &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + D \|\mathbf{y} - \mathbf{e} - \mathbf{Ax}_\eta\|_{\ell_2} \leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + D \|\mathbf{y} - \mathbf{Ax}_\eta\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2} \\ &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + (D + 1)\eta, \end{aligned} \quad (13)$$

provided that $\|\mathbf{e}\|_{\ell_2} \leq \eta$ is true. This estimate follows from combining $\|\mathbf{x}_\eta\|_{\ell_1} \leq \|\mathbf{x}\|_{\ell_1}$ and with $\|\mathbf{y} - \mathbf{Ax}_\eta\|_{\ell_2} \leq \eta$. Once more, we point out that this estimate is only valid if an appropriate η is known.

B. Nonnegative Constraints

Here we will prove now a variation of (12) (Theorem 4.25 in [12]) which holds for nonnegative vectors and matrices in \mathcal{M}^+ . For such matrices we define a condition number by

$$\kappa(\mathbf{A}) = \min\{\|\mathbf{W}\| \|\mathbf{W}^{-1}\| \mid \exists \mathbf{t} \text{ with } \mathbf{W} = \text{diag}(\mathbf{A}^T \mathbf{t}) > 0\} \quad (14)$$

Note that for diagonal matrices \mathbf{W} with non-negative entries $\kappa(\mathbf{W}) = \|\mathbf{W}\| \|\mathbf{W}^{-1}\|$.

Theorem 4. *Let $\mathbf{A} \in \mathcal{M}^+$ obeying the s -NSP with parameters ρ and τ , and let $\kappa = \kappa(\mathbf{A})$ be its condition number. If $\kappa\rho < 1$, then*

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x}) + D (\|\mathbf{t}\|_{\ell_2} + \tau) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2}$$

is true for all nonnegative vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$. The constants amount to

$$C = \kappa \frac{(1 + \kappa\rho)^2}{1 - \kappa\rho} \text{ and } D = \kappa \frac{3 + \kappa\rho}{1 - \kappa\rho}. \quad (15)$$

Comparing this to (12) reveals, that the ℓ_1 -term ($\|\mathbf{z}\|_{\ell_1} - \|\mathbf{x}\|_{\ell_1}$) has disappeared. Let us exploit this by reproducing the steps in (13). If we once more use $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, and apply the triangle inequality, we obtain

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{c_1}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + c_2 \|\mathbf{y} - \mathbf{A}\mathbf{z}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2} \quad (16)$$

This simple observation already highlights that CS-oriented algorithms, which essentially minimize the ℓ_1 -norm, are not required anymore in the non-negative case. Instead, in order to get good estimates it makes sense to minimize the r.h.s. of the bound over the “free” parameter $\mathbf{z} \geq \mathbf{0}$. Doing so, results in a *nonnegative least-squares* estimate for \mathbf{x} by minimizing $\|\mathbf{y} - \mathbf{A}\mathbf{z}\|_{\ell_2}$ subject to $\mathbf{z} \geq \mathbf{0}$. To prove this theorem, we will need two auxiliary statements.

Lemma 5. *Suppose that \mathbf{A} obeys the s -NSP with parameters ρ and τ , and set $\mathbf{W} = \text{diag}(\mathbf{w})$, where $\mathbf{w} > \mathbf{0}$ is strictly positive. Then, $\mathbf{A}\mathbf{W}^{-1}$ also obeys the s -NSP with parameters $\tilde{\rho} = \kappa(\mathbf{W})\rho$ and $\tilde{\tau} = \|\mathbf{W}\|\tau$.*

Proof: First, since \mathbf{W} is diagonal we can conclude for any vector $\mathbf{v} \in \mathbb{R}^n$ and any set $S \subset [n]$ that $\mathbf{W}^{-1}\mathbf{v}_S = (\mathbf{W}^{-1}\mathbf{v})_S$ (same for \bar{S}). Also, \mathbf{A} obeys the s -NSP which in turn implies for any $|S| \leq s$:

$$\begin{aligned} \|\mathbf{v}_S\|_{\ell_2} &= \|\mathbf{W}\mathbf{W}^{-1}\mathbf{v}_S\|_{\ell_2} \leq \|\mathbf{W}\| \|(\mathbf{W}^{-1}\mathbf{v})_S\|_{\ell_2} \leq \|\mathbf{W}\| \left(\frac{\rho}{\sqrt{s}} \|(\mathbf{W}^{-1}\mathbf{v})_{\bar{S}}\|_{\ell_2} + \tau \|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\|_{\ell_2} \right) \\ &= \tilde{\rho} \sigma_s(\mathbf{v}) + \tilde{\tau} \|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\|_{\ell_2}. \end{aligned}$$

■

Lemma 6. *Suppose that $\mathbf{W} := \text{diag}(\mathbf{A}^T \mathbf{t}) > 0$ for some $\mathbf{t} \in \mathbb{R}^m$. Then any pair $\mathbf{x}, \mathbf{z} \geq \mathbf{0}$ obeys*

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} \leq \|\mathbf{t}\|_{\ell_2} \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \quad (17)$$

Proof: Note that, by construction, \mathbf{W} is symmetric and preserves positivity of vectors. These features together with positivity of \mathbf{z} imply

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} = \langle \mathbf{1}, \mathbf{W}\mathbf{z} \rangle = \langle \mathbf{W}\mathbf{1}, \mathbf{z} \rangle = \langle \text{diag}(\mathbf{A}^T \mathbf{t}) \mathbf{1}, \mathbf{z} \rangle = \langle \mathbf{A}^T \mathbf{t}, \mathbf{z} \rangle = \langle \mathbf{t}, \mathbf{A}\mathbf{z} \rangle.$$

An analogous reformulation is true for $\|\mathbf{W}\mathbf{x}\|_{\ell_1}$ and combining these two reveals

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} = \langle \mathbf{t}, \mathbf{A}(\mathbf{z} - \mathbf{x}) \rangle \leq \|\mathbf{t}\|_{\ell_2} \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|_{\ell_2}$$

due to Cauchy-Schwarz. \blacksquare

Proof of Theorem 4: The assumption $\mathbf{A} \in \mathcal{M}^+$ assures that there exists $\mathbf{t} \in \mathbb{R}^m$ such that $\mathbf{w} = \mathbf{A}^T \mathbf{t} > \mathbf{0}$ and we define $\mathbf{W} := \text{diag}(\mathbf{w})$. By construction, \mathbf{W} is invertible and admits a condition number $\kappa = \|\mathbf{W}\| \|\mathbf{W}^{-1}\|$. Thus, we can write

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} = \|\mathbf{W}^{-1} \mathbf{W}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \leq \|\mathbf{W}^{-1}\| \|\mathbf{W}(\mathbf{x} - \mathbf{z})\|_{\ell_2}$$

for any pair $\mathbf{x}, \mathbf{z} > \mathbf{0}$. Now, since \mathbf{A} obeys the s -NSP, Lemma 5 assures that $\mathbf{A}\mathbf{W}^{-1}$ has s -NSP as well, with parameters $\tilde{\rho} = \kappa\rho$ and $\tilde{\tau} = \|\mathbf{W}\|\tau$. Thus, from (12) we conclude that for vectors $\mathbf{W}\mathbf{x}$ and $\mathbf{W}\mathbf{z}$ we have

$$\begin{aligned} \|\mathbf{W}(\mathbf{x} - \mathbf{z})\|_{\ell_2} &\leq \frac{C'}{\sqrt{s}} (\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} + 2\sigma_s(\mathbf{W}\mathbf{x})_{\ell_1}) + D' \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \\ &\stackrel{(17)}{\leq} \frac{2C' \|\mathbf{W}\|}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + \left(\frac{C' \|\mathbf{t}\|_{\ell_2}}{\sqrt{s}} + D' \right) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2}. \end{aligned}$$

Here, we invoked Lemma 6 in the last step, as well as the relation $\sigma_s(\mathbf{W}\mathbf{x})_{\ell_1} \leq \|\mathbf{W}\| \sigma_s(\mathbf{x})_{\ell_1}$. The constants above amount to $C' = \frac{(1+\tilde{\rho})^2}{1-\tilde{\rho}} = \frac{(1+\kappa\rho)^2}{1-\kappa\rho}$ and $D' = \frac{3+\tilde{\rho}}{1-\tilde{\rho}} \tilde{\tau} = \frac{3+\kappa\rho}{1-\kappa\rho} \|\mathbf{W}\|\tau$. So, in summary we obtain

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{2C' \kappa}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + \|\mathbf{W}^{-1}\| \left(\frac{C' \|\mathbf{t}\|_{\ell_2}}{\sqrt{s}} + D' \right) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2}.$$

We shall simplify the second term further by using the fact that $(1+x)^2 \leq 3+x$ for any $x \in [0, 1]$, i.e., C' and D'/τ are both upper bounded by $\frac{3+\kappa\rho}{1-\kappa\rho} \|\mathbf{W}\|$. Consequently,

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq 2 \frac{\kappa C'}{\sqrt{s}} \sigma_s(\mathbf{x}) + \frac{3+\kappa\rho}{1-\kappa\rho} \kappa (\|\mathbf{t}\|_{\ell_2} + \tau) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2},$$

and setting $C := \kappa C'$ and $D = \frac{3+\kappa\rho}{1-\kappa\rho} \kappa$ proves the claim. \blacksquare

IV. ROBUST NSP FOR 0/1-BERNOULLI MATRICES

In this section, we prove our second main result, namely Theorem 2. Said statements summarizes two results, namely (i) 0/1-Bernoulli matrices \mathbf{A} with $m = Cs \log(n)$ rows obey the robust null space property of order s and (ii) the row space of \mathbf{A}^T allows for constructing a strictly positive vector $\mathbf{w} = \mathbf{A}^T \mathbf{t} > \mathbf{0}$ (that is sufficiently well-conditioned). We will first state the main ideas and prove both statements in subsequent subsections.

A. Sampling model and overview of main proof ideas

Let us start by formally defining the concept of a 0/1-Bernoulli matrix.

Definition 7. We call $\mathbf{A} \in \mathbb{R}^{m \times n}$ a 0/1-Bernoulli matrix with parameter $p \in [0, 1]$, if every matrix element $[\mathbf{A}]_{i,j}$ is an independent realization of a Bernoulli random variable b with parameter p , i.e.

$$\Pr[b = 1] = p \quad \text{and} \quad \Pr[b = 0] = 1 - p.$$

Recall that such a Bernoulli variable obeys $\mathbb{E}[b] = p$ and $\text{Var}(b) = \mathbb{E}[(b - \mathbb{E}[b])^2] = p(1 - p)$. By construction, the m rows $\mathbf{a}_1, \dots, \mathbf{a}_m$ of such a Bernoulli matrix are independent and obey

$$\mathbb{E}[\mathbf{a}_k] = \sum_{j=1}^n \mathbb{E}[\mathbf{A}_{k,j}] \mathbf{e}_j = p \sum_{j=1}^n \mathbf{e}_j = p\mathbf{1}.$$

This expected behavior of the individual rows will be crucial for addressing the second point in Theorem 2: setting

$$\mathbf{w} := \frac{1}{pm} \sum_{k=1}^m \mathbf{a}_k = \mathbf{A}^T \left(\frac{1}{pm} \mathbf{1} \right)$$

results in a random vector $\mathbf{w} \in \mathbb{R}^n$ that obeys $\mathbb{E}[\mathbf{w}] = \mathbf{1} > \mathbf{0}$. Applying a large deviation bound will in turn imply that a realization of \mathbf{w} will w.h.p. not deviate too much from its expectation $\mathbf{1}$ and thus remains strictly positive.

We will do this in Subsection IV-C.

However, when turning our focus to establishing null space properties for \mathbf{A} , working with 0/1-Bernoulli entries renders such a task more challenging. The simple reason for such a complication is that the individual random entries of \mathbf{A} are not centered, i.e. $\mathbb{E}[\mathbf{A}_{k,j}] = p \neq 0$. Combining this with independence of the individual entries yields

$$\mathbb{E}[\mathbf{a}_k \mathbf{a}_k^T] = p^2 \mathbf{1} \mathbf{1}^T + p(1 - p) \mathbb{I}.$$

This matrix admits a condition number of $\kappa(\mathbb{E}[\mathbf{a}_k \mathbf{a}_k^T]) = 1 + \frac{pn}{1-p}$ which underlines the ensemble's anisotropy. Traditional proof techniques, e.g. establishing an RIP, are either not applicable in such a setting, or yield sub-optimal results [13], [14]. This is not true for Mendelson's small ball method [15], [16] (see also [17]) – a strong general purpose tool whose applicability only requires row-wise independence. In the conceptually similar problem of reconstructing low rank matrix from rank-one projective measurements (which arises e.g. from the PhaseLift approach for phase retrieval [28], [18]) applying this technique allowed for establishing strong null space properties, despite a similar degree of anisotropy in the sampling model [19]. In the next subsection, we adapt the ideas from said paper to our Bernoulli model and succeed in establishing the NSP presented in Theorem 2.

Finally, we point out that a detailed survey of the applicability of Mendelson's small ball method for compressed sensing was recently presented in [20]. However, there centeredness of the individual matrix entries is a key assumption which is not met in our 0/1-Bernoulli model.

B. Null Space Properties for 0/1-Bernoulli matrices

Recall that Definition 3 states that a $m \times n$ matrix \mathbf{A} obeys the robust null space property with parameters $\rho \in (0, 1)$ and $\tau > 0$, if

$$\|\mathbf{v}_S\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_S\|_{\ell_1} + \tau \|\mathbf{A}\mathbf{v}\|_{\ell_2} \quad (18)$$

is true for all vectors $\mathbf{v} \in \mathbb{R}^n$ and support sets $S \in [n]$ with support size $|S| \leq s$. Demanding such generality in the choice of the support set is in fact not necessary, see e.g. [12, Remark 4.2]. For a fixed vector \mathbf{v} , the above condition holds for any index set S , if it holds for an index set S_{\max} containing the s largest (in modulus) entries

of \mathbf{v} . Introducing the notation $\mathbf{v}_s := \mathbf{v}_{S_{\max}}$ and $\mathbf{v}_c := \mathbf{v}_{\bar{S}_{\max}}$, the robust null space property (18) holds, provided that every vector $\mathbf{v} \in \mathbb{R}^n$ obeys

$$\|\mathbf{v}_s\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1} + \tau \|\mathbf{A}\mathbf{v}\|_{\ell_2}. \quad (19)$$

Note that this requirement is invariant under re-scaling and we may w.l.o.g. assume $\|\mathbf{v}\|_{\ell_2} = 1$. Moreover, for fixed parameters s and ρ , any vector \mathbf{v} obeying $\|\mathbf{v}_s\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1}$ is guaranteed to fulfill (19) by default. Consequently, when aiming to establish null space properties, it suffices to establish condition (19) for the set of unit-norm vectors that do not obey this criterion:

$$T_{\rho,s} := \left\{ \mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\|_{\ell_2} = 1, \|\mathbf{v}_s\|_{\ell_2} > \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1} \right\}.$$

As a result, a matrix \mathbf{A} obeys the NSP (18), if

$$\inf \{ \|\mathbf{A}\mathbf{v}\|_{\ell_2} : \mathbf{v} \in T_{\rho,s} \} > \frac{1}{\tau}, \quad (20)$$

holds, where $\tau > 0$ is the second parameter appearing in (18). The task of establishing this is somewhat simplified by the observation that the set $T_{\rho,s}$ exclusively contains vectors that are effectively sparse:

Lemma 8. *For fixed s and ρ , every vector $\mathbf{v} \in T_{\rho,s}$ obeys*

$$\|\mathbf{v}\|_{\ell_1} \leq \sqrt{s} \frac{1 + \rho}{\rho} \|\mathbf{v}\|_{\ell_2}.$$

Proof: Note that any \mathbf{v}_s is s -sparse by construction and thus obeys $\|\mathbf{v}_s\|_{\ell_1} \leq \sqrt{s} \|\mathbf{v}_s\|_{\ell_2}$. Combining this with the triangle inequality and the defining feature of the set $T_{\rho,s}$ yields

$$\|\mathbf{v}\|_{\ell_1} = \|\mathbf{v}_s + \mathbf{v}_c\|_{\ell_1} \leq \|\mathbf{v}_s\|_{\ell_1} + \|\mathbf{v}_c\|_{\ell_1} \leq \sqrt{s} \|\mathbf{v}_s\|_{\ell_2} + \frac{\sqrt{s}}{\rho} \|\mathbf{v}_s\|_{\ell_2}$$

and the claim readily follows from $\|\mathbf{v}_s\|_{\ell_2} \leq \|\mathbf{v}\|_{\ell_2}$. ■

Despite such a geometric insight, proving (20) for a given \mathbf{A} is still a daunting task. This situation greatly changes, if we assume that our sampling matrix \mathbf{A} consists of m rows $\mathbf{a}_1, \dots, \mathbf{a}_m$ that are independent instances of a random vector $\mathbf{a} \in \mathbb{R}^n$. Assuming this, (20) is equivalent to showing

$$\inf_{\mathbf{v} \in T_{\rho,s}} \left(\sum_{k=1}^m |\langle \mathbf{a}_k, \mathbf{v} \rangle|^2 \right)^{1/2} > \frac{1}{\tau}. \quad (21)$$

Independence of the \mathbf{a}_k 's then allows for establishing this (w.h.p.) by resorting to Mendelson's small ball method [15], [16], [17]:

Theorem 9 (Koltchinskii, Mendelson; Tropp's version [17]). *Fix $E \subset \mathbb{R}^n$ and let $\mathbf{a}_1, \dots, \mathbf{a}_m$ be independent copies of a random vector $\mathbf{a} \in \mathbb{R}^n$. Set $\mathbf{h} = \frac{1}{\sqrt{m}} \sum_{k=1}^m \epsilon_k \mathbf{a}_k$, where $\epsilon_1, \dots, \epsilon_m$ is a Rademacher sequence, and for $\xi > 0$ define*

$$Q_\xi(E, \mathbf{a}) = \inf_{\mathbf{u} \in E} \Pr [|\langle \mathbf{a}, \mathbf{u} \rangle| \geq \xi], \quad \text{as well as} \quad W_m(E, \mathbf{a}) = \mathbb{E} \left[\sup_{\mathbf{u} \in E} \langle \mathbf{h}, \mathbf{u} \rangle \right].$$

Then, for any $\xi > 0$ and $t \geq 0$ the following is true with probability at least $1 - e^{-2t^2}$:

$$\inf_{\mathbf{u} \in E} \left(\sum_{k=1}^m |\langle \mathbf{a}_k, \mathbf{u} \rangle|^2 \right)^{1/2} \geq \xi \sqrt{m} Q_{2\xi}(E, \mathbf{a}) - \xi t - 2W_m(E, \mathbf{a}). \quad (22)$$

In order to establish (21), we can set $E = T_{\rho,s}$, choose ξ and t appropriately and establish suitable bounds for $Q_\xi(T_{\rho,s}, \mathbf{a})$ and $W_m(T_{\rho,s}, \mathbf{a})$. Note that the geometric insight provided in Lemma 8 considerably simplifies this last task. It assures

$$W_m(T_{\rho,s}, \mathbf{a}) = \mathbb{E} \left[\sup_{\mathbf{u} \in T_{\rho,s}} \langle \mathbf{h}, \mathbf{u} \rangle \right] \leq \sup_{\mathbf{u} \in T_{\rho,s}} \|\mathbf{u}\|_{\ell_1} \mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{s} \frac{1+\rho}{\rho} \mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}]$$

and it suffices bound $\mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}]$ from above. We do this by adapting the techniques from [29, Proposition 13] to the vector case. The calculations are detailed in the appendix and yield

$$\mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{4p(1-p) \left(3 \log(2n) + \frac{p}{1-p} \right)} \quad (23)$$

under the assumption that the sampling rate m exceeds $\frac{\log(n)}{p^2(1-p)^2}$. Such a bound allows us to deduce

$$W_m(T_{\rho,s}, \mathbf{a}) \leq \frac{1+\rho}{\rho} \sqrt{4sp(1-p) \left(3 \log(2n) + \frac{p}{1-p} \right)} \quad (24)$$

without having to pay too much attention to the complicated geometry of the set $T_{\rho,s}$. Likewise, said set is strictly contained in the unit sphere $S^{n-1} \in \mathbb{R}^n$. For fixed $\xi > 0$, this allows us to bound $Q_{2\xi}(T_{\rho,s}, \mathbf{a})$ from below by establishing a global lower bound on $\Pr[|\langle \mathbf{a}, \mathbf{u} \rangle| \geq 2\xi]$ that is valid for any $\mathbf{u} \in S^{n-1}$. We do this in the appendix and obtain

$$\Pr \left[|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)} \right] \geq \frac{4}{13} p(1-p)(1-\theta^2)^2 \quad \forall \mathbf{z} \in S^{n-1} \text{ and } \theta \in [0, 1].$$

The structure of such a global bound suggests choosing $\xi = \frac{1}{4} \sqrt{p(1-p)}$ for which we can conclude

$$Q_{2\xi}(T_{\rho,s}, \mathbf{a}) \geq \frac{4p(1-p)(\frac{3}{4})^2}{13} > \frac{p(1-p)}{6}. \quad (25)$$

Such a choice of ξ , setting $t = \frac{p(1-p)}{12} \sqrt{m}$ and evoking the bounds (24) and (25) into (22) implies

$$\begin{aligned} \inf_{\mathbf{v} \in T_{\rho,s}} \|\mathbf{A}\mathbf{v}\|_{\ell_2} &\geq \frac{\sqrt{p(1-p)}^3}{24} \sqrt{m} - \frac{\sqrt{p(1-p)}^3}{48} \sqrt{m} - 2 \frac{1+\rho}{\rho} \sqrt{4sp(1-p) \left(3 \log(2n) + \frac{p}{1-p} \right)} \\ &= \sqrt{p(1-p)} \left(\frac{p(1-p)}{48} \sqrt{m} - \sqrt{16 \frac{(1+\rho)^2}{\rho^2} s \left(3 \log(2n) + \frac{p}{1-p} \right)} \right) \end{aligned}$$

with probability at least $1 - e^{-\frac{p^2(1-p)^2}{72} m}$. This prompts us to demand

$$m \geq \frac{C_1(1+\rho)^2}{p^2(1-p)^2 \rho^2} s \left(\log(n) + \frac{p}{1-p} \right), \quad (26)$$

where C_2 is a sufficiently large constant (note that this justifies the assumption $m \geq \frac{\log(n)}{p^2(1-p)^2}$ made before). Then the above inequality implies that there is another constant $C_2 > 0$ (whose size depends on the choice of C) such that

$$\inf_{\mathbf{v} \in T_{\rho,s}} \|\mathbf{A}\mathbf{v}\|_{\ell_2} \geq \frac{\sqrt{p(1-p)}^3}{C_2} \sqrt{m}. \quad (27)$$

with probability of failure bounded by $e^{-\frac{p^2(1-p)^2}{72}m}$. Comparing this bound to (21) allows us to set $\tau = \frac{C_2}{\sqrt{p(1-p)^3}\sqrt{m}}$ and we arrive at the main result of this section:

Theorem 10. *Let $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a 0/1-Bernoulli matrix with parameter $p \in [0, 1]$ and fix $s \leq n$ and $\rho \in [0, 1]$. Then, there are constants C_1 and C_2 such that choosing the number of rows to be*

$$m = C_1 \frac{(1+\rho)^2}{p^2(1-p)^2\rho^2} s \left(\log(n) + \frac{p}{1-p} \right) \quad (28)$$

assures that \mathbf{A} obeys the robust NSP of order s with parameters ρ and $\tau = \frac{C_2}{\sqrt{p(1-p)^3}\sqrt{m}}$. Hereby, the probability of failure is bounded by $e^{-\frac{p^2(1-p)^2}{72}m}$.

This is a more detailed version of the first claim presented in Theorem 2. We see that sampling rate, size of the NSP-parameter τ and the probability bound all depend on the Bernoulli parameter $p \in [0, 1]$. Factoring out the p -dependence of m by writing $m = \frac{\tilde{m}}{p^2(1-p)^2}$ we obtain a probability bound of $e^{-\frac{\tilde{m}}{72}}$ which is independent of p . On the other hand $\tau = \frac{C_2}{\sqrt{p(1-p)\tilde{m}}}$ still exhibits a p -dependence.

Finally, we point out that when opting for a standard Bernoulli process, i.e. $p = \frac{1}{2}$, the assertions of Theorem 10 considerably simplify, because $p(1-p) = \frac{1}{4}$. Inserting this, we obtain:

Corollary 11. *Fix $s \leq n$, $\rho \in [0, 1]$ and let \mathbf{A} be a standard $(m \times n)$ 0/1-Bernoulli matrix (i.e. $p = \frac{1}{2}$) with*

$$m \geq 17C_1 \frac{(1+\rho)^2}{\rho^2} s \log(n).$$

Then with probability at least $1 - e^{-\frac{m}{1152}}$ this matrix obeys the NSP of order s with parameters ρ and $\tau = \frac{C_2}{2\sqrt{m}}$. Hence, C_1 and C_2 are the constants from Theorem 10.

C. 0/1-Bernoulli matrices lie in \mathcal{M}_+

We now move on to showing that 0/1-Bernoulli matrices are very likely to admit the second requirement of Theorem 4. Namely, that there exists a vector $\mathbf{w} = \mathbf{A}^T \mathbf{t}$ that is strictly positive which is equivalent to demanding $\mathbf{A} \in \mathcal{M}_+$. Concretely, we show that setting $\mathbf{t} = \frac{1}{pm} \mathbf{1} \in \mathbb{R}^m$ results in a strictly positive vector $\mathbf{w} \in \mathbb{R}^n$ whose conditioning obeys

$$\kappa(\mathbf{w}) = \frac{\max_k |\langle \mathbf{e}_k, \mathbf{w} \rangle|}{\min_k |\langle \mathbf{e}_k, \mathbf{w} \rangle|} \leq 3. \quad (29)$$

To do so, we note that $\mathbf{w} = \frac{1}{pm} \sum_{k=1}^m \mathbf{a}_k$ has expectation $\mathbb{E}[\mathbf{w}] = \mathbf{1}$, which is – up to re-scaling – the unique non-negative vector admitting $\kappa(\mathbf{1}) = 1$. After having realized this, it suffices to use a concentration inequality to prove that w.h.p. \mathbf{w} does not deviate too much from its expectation $\mathbf{1}$. We do this by invoking a Bernstein inequality which implies:

Theorem 12. *Suppose that $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a 0/1-Bernoulli matrix with parameter $p \in [0, 1]$ and set*

$$\mathbf{w} = \mathbf{A}^T \mathbf{t} \in \mathbb{R}^n \quad \text{with} \quad \mathbf{t} = \frac{1}{pm} \mathbf{1} \in \mathbb{R}^m. \quad (30)$$

Then with probability at least $1 - ne^{-\frac{3}{8}p(1-p)m}$ $\max_i |\langle \mathbf{e}_i, \mathbf{w} \rangle| \leq \frac{3}{2}$ and $\min_i |\langle \mathbf{e}_i, \mathbf{w} \rangle| \geq \frac{1}{2}$ which in turn implies (29).

Proof: Instead of showing the claim directly, we prove that a stronger statement, namely

$$|\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| \leq \frac{1}{2} \quad 1 \leq i \leq n, \quad (31)$$

is true with probability of failure bounded by $ne^{-\frac{3}{8}p(1-p)m}$. If such a bound is true for all i , it is also valid for maximal and minimal components and we obtain

$$\max_i |\langle \mathbf{e}_i, \mathbf{w} \rangle| \leq \max_k |\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| + 1 \leq \frac{3}{2} \quad \text{and} \quad \min_k |\langle \mathbf{e}_i, \mathbf{w} \rangle| \geq 1 - \max_i |\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| \geq \frac{1}{2},$$

as claimed. In order to prove (31), we fix $1 \leq i \leq n$ and focus on

$$|\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| = \left| \frac{1}{pm} \sum_{k=1}^m \langle \mathbf{e}_i, \mathbf{a}_k \rangle - 1 \right| = \frac{1}{pm} \left| \sum_{k=1}^m (b_{k,i} - \mathbb{E}[b_{k,i}]) \right|.$$

Here, we have used $\langle \mathbf{e}_i, \mathbf{a}_k \rangle = \langle \mathbf{e}_k, \mathbf{A} \mathbf{e}_i \rangle = b_{k,i}$, which is an independent instance of a Bernoulli random variable with parameter p . Thus we are faced with bounding the deviation of a sum of m centered, independent random variables $c_k := b_{k,i} - \mathbb{E}[b_{k,i}]$ from its mean. Each such variable obeys

$$|c_k| \leq \max\{p, 1-p\} \leq 1 \quad \text{and} \quad \mathbb{E}[c_k^2] = \text{Var}(b_{k,i}) = p(1-p).$$

Applying a Bernstein inequality [12, Theorem 7.30] reveals

$$\Pr \left[|\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| \geq \frac{1}{2} \right] \leq \Pr \left[|\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| \geq \frac{1-p}{2} \right] = \Pr \left[\left| \sum_{k=1}^m c_k \right| \geq \frac{mp(1-p)}{2} \right] \leq \exp \left(-\frac{3}{8}p(1-p)m \right).$$

Combining this with a union bound assures that $|\langle \mathbf{e}_i, \mathbf{w} \rangle - 1| < \frac{1}{2}$ is simultaneously true for all $1 \leq i \leq n$ with probability at least $1 - ne^{-\frac{3}{8}p(1-p)m}$. ■

D. Proof of Theorem 2

Finally, these two results can be combined to yield Theorem 2. It readily follows from taking a union bound over the individual probabilities of failure. Theorem 10 requires a sampling rate of

$$m \geq C_1 \frac{(1+\rho)^2}{p^2(1-p)^2 \rho^2 s} \left(\log(n) + \frac{p}{1-p} \right) \quad (32)$$

to assure that a corresponding 0/1-Bernoulli matrix obeys a strong version of the NSP with probability at least $1 - e^{-\frac{p^2(1-p)^2}{72}m}$. On the other hand, Theorem 12 asserts that choosing $\mathbf{w} = \mathbf{A}^T \frac{1}{pm} \mathbf{1}$ for 0/1-Bernoulli matrices \mathbf{A} results in a well-conditioned and strictly positive vector \mathbf{w} with probability at least $1 - ne^{-\frac{3}{8}p(1-p)m}$. The probability that either of these assertions fails to hold can be controlled by the union bound over both probabilities of failure:

$$\begin{aligned} \Pr [\text{Thm. 10 fails to hold} \cup \text{Thm. 12 fails to hold}] &\leq \Pr [\text{Thm. 10 fails to hold}] + \Pr [\text{Thm. 12 fails to hold}] \\ &\leq e^{-\frac{p^2(1-p)^2}{72}m} + ne^{-\frac{3p(1-p)}{8}m} \leq (n+1)e^{-\frac{p^2(1-p)^2}{72}m}. \end{aligned}$$

Finally, we focus on 0/1-Bernoulli matrices \mathbf{A} for which both statements are true and whose sampling rate exceeds (32). Theorem 10 then implies that \mathbf{A} obeys the s -NSP with a pre-selected parameter $\rho \in [0, 1]$ and

$\tau = \frac{C_2}{\sqrt{p(1-p)}^3 \sqrt{m}}$. Moreover, the vector selection $\mathbf{t} = \frac{1}{pm} \mathbf{1}$ in Theorem 12 obeys $\|\mathbf{t}\|_{\ell_2} = \frac{1}{p\sqrt{m}}$. As a result, the implication of Theorem 4 reads for any $\mathbf{x}, \mathbf{z} \geq \mathbf{0}$:

$$\begin{aligned} \|\mathbf{x} - \mathbf{z}\|_{\ell_2} &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x}) + D (\|\mathbf{t}\|_{\ell_2} + \tau) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \\ &= \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x}) + D \left(\frac{1}{p\sqrt{m}} + \frac{C_2}{\sqrt{p(1-p)}^3 \sqrt{m}} \right) \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \\ &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x}) + \frac{D(1+C_2)}{\sqrt{p(1-p)}^3} \frac{\|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2}}{\sqrt{m}}. \end{aligned}$$

The constant $\frac{D(1+C_2)}{\sqrt{p(1-p)}^3}$ is the explicit value of D' in Theorem 1 for the case of 0/1-Bernoulli matrices with parameter $p \in [0, 1]$.

V. NUMERICAL EXPERIMENTS

In the following we evaluate the *nonnegative least squares* (NNLS) in (3) and we compare this to the results obtained with *basis pursuit denoising* (BPDN) in (10). The NNLS has been computed using the `lsqnonneg` function in MATLAB which implements the “active-set” Lawson–Hanson algorithm [30]. For the BPDN the SPGL1 toolbox has been used [31].

In a first test we have evaluated numerically the phase transition of NNLS in the 0/1-Bernoulli setting for the noiseless case. The dimension and sparsity parameters are generated uniformly (in this order) in the ranges $n \in [10 \dots 500]$, $m \in [10 \dots n]$ and $s \in [1 \dots m]$. Thus, the sparsity/density variable is $\rho = s/m$ and the subsampling ratio is $\delta = m/n$. The $m \times n$ measurement matrix \mathbf{A} is generated using the iid. 0/1-Bernoulli model with $p = 1/2$. The nonnegative s -sparse signal $0 \leq \mathbf{x} \in \mathbb{R}^n$ to recover is created as follows: the random support $\text{supp}(\mathbf{x})$ is obtained from taking the first s elements of a random (uniformly-distributed) permutation of the indices $(1 \dots n)$. On this support each value is the absolute value of an iid. standard (zero mean, unit variance) Gaussian, i.e., $x_i = |g_i|$ with $g_i \sim N(0, 1)$ for all $i \in \text{supp}(\mathbf{x})$. An event counts as successful once $\|\mathbf{x} - \hat{\mathbf{x}}\|_{\ell_2} \leq 10^{-3} \|\mathbf{x}\|_{\ell_2}$. The resulting phase transition diagram, shown in Figure 1 above, demonstrates that NNLS indeed reliably recovers nonnegative sparse vectors without any ℓ_1 -regularization.

In the second experiment we consider the noisy case. Beside its simplicity, the important feature of NNLS is that no a-priori norm assumptions on the noise are necessary as it is required for the BPDN. As illustrated in (4), a result of Theorem 1 is that the NNLS estimate $\mathbf{x}^\#$ fulfils:

$$\|\mathbf{x} - \mathbf{x}^\#\|_{\ell_2} \leq \frac{2C}{\sqrt{m}} \|\mathbf{e}\|_{\ell_2} \quad (33)$$

A similar bound is valid for the BPDN (see (13)) estimate \mathbf{x}_η when $\|\mathbf{e}\|_{\ell_2} \leq \eta$, i.e., once $\|\mathbf{e}\|_{\ell_2}$ is known. Interestingly, even under this prerequisites the performance of NNLS is considerable better than BPDN in our setting. This is visualized in Figure 2 where each component e_j of \mathbf{e} is iid. Gaussian distributed with zero mean and variance $\sigma_e^2 = 1/100$. There recovery has been identified as “successful” if (33) is fulfilled for $2C = \sqrt{10}$.

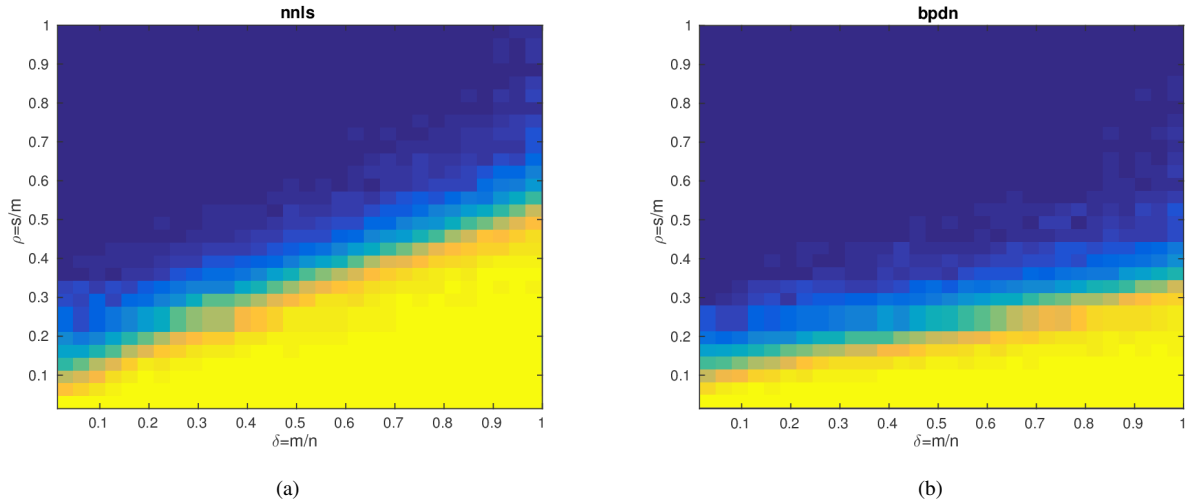


Fig. 2: Comparison of NNLS in (3) with BPDN in (10) for iid. 0/1-Bernoulli matrices in the noisy setting.

VI. CONCLUSIONS

In this work we have shown that nonnegativity is a tremendous important additional property when recovering sparse vectors. This situation is relevant in many applications and we are motivated here by activity detection in wireless networks using individual sequences. Designing measurement matrices such that convex hull of its columns (the sequences) is sufficiently well-separated from the origin recovery allow for remarkable simple recovery algorithms which are prone to noise and blind in a sense that no regularization and a priori information on the noise is required. We have demonstrated this feature by strengthen the implications of the robust nullspace property for the nonnegative setting. Furthermore, we have shown that iid. binary measurements fulfill w.h.p. this property and are simultaneously well-conditioned and can be used therefore for recovering nonnegative and sparse vectors in the optimal regime.

ACKNOWLEDGEMENTS

The authors want to thank S. Dirksen, H. Rauhut and D. Gross for inspiring discussions and helpful comments. This work has been supported by “Mathematics of Signal Processing” trimester program at the Hausdorff Research Institute for Mathematics (HIM). RK acknowledges support from the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contract W911NF-14-1-0098 (Quantum Characterization, Verification, and Validation), and the DFG. PJ is supported by DFG grant JU 2795/2.

APPENDIX: PROOFS OF EQUATIONS (23) AND (25)

Here we provide proofs of the two bounds (23) and (25) on which we built our argument that 0/1-Bernoulli matrices obey the robust NSP. Since both are rather technical and not essential for understanding the main ideas, we decided to present them in this appendix.

Preliminaries

In order to prove the remaining estimates we rely on a couple of probabilistic standard tools which we shall summarize here. Recall that a Rademacher sequence $(\epsilon_1, \dots, \epsilon_m)$ is a sequence of m independent dichotomic random variables obeying $\Pr[\epsilon_k = 1] = \Pr[\epsilon_k = -1] = \frac{1}{2}$.

Theorem 13 (Khinchine Inequality, Corollary 8.7 in [12]). *Let $\mathbf{c} \in \mathbb{C}^m$ and $\epsilon_1, \dots, \epsilon_m$ be a Rademacher sequence. Then for all $q > 0$*

$$\left(\mathbb{E} \left[\left| \sum_{k=1}^m \epsilon_k c_k \right|^q \right] \right)^{1/q} \leq 2^{3/(4q)} e^{-1/2} \|\mathbf{c}\|_{\ell_2}.$$

Theorem 14 (Non-commutative Khinchine inequality: Exercise 8.6 (d) in [29]). *Let M_1, \dots, M_m be hermitian $n \times n$ matrices and suppose that $(\epsilon_1, \dots, \epsilon_m)$ is a Rademacher sequence. Then*

$$\mathbb{E} \left[\left\| \sum_{k=1}^m \epsilon_k M_k \right\|^2 \right] \leq \sqrt{2 \log(2n)} \left\| \sum_{k=1}^m M_k^2 \right\|^{1/2}.$$

Theorem 15 (Matrix Chernoff for expectation values: Theorem 5.1.1 in [32] (see also [33])). *Let $\{M_k\}_{k=1}^m$ be a sequence of independent, random, non-negative $n \times n$ matrices obeying $\|M_k\| \leq R$ almost surely. Then, for any $t > 0$ their sum obeys*

$$\mathbb{E} \left[\left\| \sum_{k=1}^m M_k \right\|^2 \right] \leq \frac{e^t - 1}{t} \left\| \sum_{k=1}^m \mathbb{E}[M_k] \right\|_{\infty} + \frac{R}{t} \log(n).$$

Theorem 16 (Paley-Zygmund Inequality). *Let X be a non-negative random variable with bounded second moment. Then*

$$\Pr[X \geq \theta \mathbb{E}[X]] \geq \frac{(1 - \theta)^2 \mathbb{E}[X]}{\text{Var}(X) + \mathbb{E}[X]^2},$$

where $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ is the variance of X .

Bounding $\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}]$ for 0/1-Bernoulli matrices

In this section, we prove that the bound presented in (23) holds in the Bernoulli setting. Let \mathbf{A} be a 0/1-Bernoulli matrix with parameter p and m rows $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$. The vector $\mathbf{h} := \frac{1}{\sqrt{m}} \sum_{k=1}^m \epsilon_k \mathbf{a}_k$ was introduced in Theorem 9 and in (23) we claimed that this vector obeys

$$\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{4p(1-p) \left(3 \log(2n) + \frac{p}{1-p} \right)}, \quad (34)$$

provided that $m \geq \frac{\log(n)}{p^2(1-p)^2}$. When aiming to prove this, we first minimize the anisotropic impact of \mathbf{A} 's rows. Recalling $\mathbb{E}[\mathbf{a}_k] = p\mathbf{1}$, we introduce $\tilde{\mathbf{a}}_k := \mathbf{a}_k - p\mathbf{1}$, and likewise $\tilde{\mathbf{h}} := \frac{1}{\sqrt{m}} \sum_{k=1}^m \epsilon_k \tilde{\mathbf{a}}_k$, which obey

$$\mathbf{h} = \tilde{\mathbf{h}} + \frac{p}{\sqrt{m}} \left(\sum_{k=1}^m \epsilon_k \right) \mathbf{1} \quad (35)$$

by construction. Applying the triangle inequality reveals

$$\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}] \leq \mathbb{E}[\|\tilde{\mathbf{h}}\|_{\ell_\infty}] + \frac{p}{\sqrt{m}} \mathbb{E} \left[\left\| \sum_{k=1}^m \epsilon_k \right\| \right] \|\mathbf{1}\|_{\ell_\infty} \quad (36)$$

and we may bound the two terms individually. For the second term, we resort to the classical Khintchine inequality (with $q = 1$ and $c = 1$) and obtain

$$\frac{p}{\sqrt{m}} \left(\sum_{k=1}^m \epsilon_k \right) \mathbf{1} \leq \frac{p 2^{3/4} e^{-1/2}}{\sqrt{m}} \|\mathbf{1}\|_{\ell_2} \|\mathbf{1}\|_{\ell_\infty} \leq \sqrt{2} p, \quad (37)$$

because $\|\mathbf{1}\|_{\ell_2} = \sqrt{m} \|\mathbf{1}\|_{\ell_\infty} = \sqrt{m}$ and $2^{3/4} e^{-1/2} \simeq 1.02.006 < \sqrt{2}$. For the remaining estimate of $\mathbb{E} \left[\|\tilde{\mathbf{h}}\|_{\ell_\infty} \right]$, we find it advantageous to work with an equivalent matrix problem

$$\mathbb{E} \left[\|\tilde{\mathbf{h}}\|_{\ell_\infty} \right] = \mathbb{E} \left[\left\| \text{diag}(\tilde{\mathbf{h}}) \right\|_{\ell_\infty} \right] = \frac{1}{\sqrt{m}} \mathbb{E} \left[\left\| \sum_{k=1}^m \epsilon_k \text{diag}(\tilde{\mathbf{a}}_k) \right\|_{\ell_\infty} \right]$$

that can be tackled by consecutively applying matrix Khintchine and Chernoff inequalities. Exploiting the randomness in $(\epsilon_1, \dots, \epsilon_m)$, by applying Theorem 14 assures

$$\begin{aligned} \mathbb{E} \left[\|\tilde{\mathbf{h}}\|_{\ell_\infty} \right] &= \frac{1}{\sqrt{m}} \mathbb{E}_{\mathbf{a}} \mathbb{E}_{\epsilon} \left[\left\| \sum_{k=1}^m \epsilon_k \text{diag}(\tilde{\mathbf{a}}_k) \right\|_{\ell_\infty} \right] \leq \sqrt{\frac{2 \log(2n)}{m}} \mathbb{E}_{\mathbf{a}} \left[\left\| \sum_{k=1}^m \text{diag}(\tilde{\mathbf{a}}_k)^2 \right\|_{\ell_\infty}^{1/2} \right] \\ &\leq \sqrt{\frac{2 \log(2n)}{m}} \left(\mathbb{E}_{\mathbf{a}} \left[\left\| \sum_{k=1}^m \text{diag}(\tilde{\mathbf{a}}_k)^2 \right\|_{\ell_\infty} \right] \right)^{1/2}, \end{aligned} \quad (38)$$

where we have also employed Jensen's inequality. Now, note that the matrices $\text{diag}(\tilde{\mathbf{a}}_k)^2$ are all positive semidefinite and obey

$$\begin{aligned} \left\| \text{diag}(\tilde{\mathbf{a}}_k)^2 \right\| &= \left\| \text{diag}(\mathbf{a}_k - p\mathbf{1})^2 \right\| \leq \max \{p^2, (1-p)^2\}, \\ \mathbb{E} \left[\text{diag}(\tilde{\mathbf{a}}_k)^2 \right] &= \sum_{i=1}^n \mathbb{E} \left[(\langle \mathbf{e}_i, \mathbf{a}_k \rangle - p)^2 \right] \mathbf{e}_i \mathbf{e}_i^T = p(1-p) \mathbb{I}. \end{aligned}$$

This is true, because each $\langle \mathbf{e}_i, \mathbf{a}_k \rangle$ is an independent instance of a Bernoulli variable with parameter p . Thus, Theorem 15 is applicable and setting $t = 1$ implies for

$$\begin{aligned} \mathbb{E}_{\mathbf{a}} \left[\left\| \sum_{k=1}^m \text{diag}(\tilde{\mathbf{a}}_k)^2 \right\|_{\ell_\infty} \right] &\leq (e-1) \left\| \sum_{k=1}^m p(1-p) \mathbb{I} \right\|_{\ell_\infty} + \max \{p^2, (1-p)^2\} \log(n) \\ &\leq ep(1-p)m + \max \{p^2, (1-p)^2\} \log(n). \end{aligned}$$

Inserting this into (38) yields

$$\mathbb{E} \left[\|\tilde{\mathbf{h}}\|_{\ell_\infty} \right] \leq \sqrt{2 \log(2n) \left(ep(1-p) + \frac{\log(n)}{m} \right)} \quad (39)$$

and turning back to (36), we see that

$$\mathbb{E} \left[\|\mathbf{h}\|_{\ell_\infty} \right] \leq \sqrt{2 \log(2n) \left(ep(1-p) + \frac{\log(n)}{m} \right)} + \sqrt{2} p$$

holds. In order to simplify this further, we now use the prior assumption $m \geq \frac{\log(n)}{p^2(1-p)^2}$ which assures

$$\frac{\log(n)}{m} \leq p^2(1-p)^2 \leq \frac{1}{4} p(1-p),$$

because $p(1-p) \leq \frac{1}{4}$ for all $p \in [0, 1]$. Combining this with $e + \frac{1}{4} < 3$ allows us to deduce

$$\mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{6p(1-p) \log(2n)} + \sqrt{2}p.$$

Finally, we use the elementary inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)} \forall a, b \geq 0$ to obtain

$$\mathbb{E} [\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{2(6p(1-p) \log(2n) + 2p^2)} = \sqrt{4p(1-p) \left(3 \log(2n) + \frac{p}{1-p}\right)},$$

which is the estimate presented in (34).

A. Bounding $\Pr [|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \|\mathbf{z}\|_{\ell_2}]$ for 0/1-Bernoulli vectors

In this final section we prove that for any unit vector $\mathbf{z} = (z_1, \dots, z_n)^T \in S^{n-1}$ and any $\theta \in [0, 1/2]$, the bound

$$\Pr [|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)}] \geq \frac{4}{13} p(1-p)(1-\theta^2)^2 \quad (40)$$

holds in the Bernoulli setting. Here, the probability is taken over instances $\mathbf{a} \in \mathbb{R}^n$ of the i.i.d. row distribution in a 0/1-Bernoulli matrix. Hence, $\mathbf{a} = \sum_{i=1}^n b_i \mathbf{e}_i$, where each b_i is an independent Bernoulli random variable with parameter p . This estimate is going to rely on the Paley-Zygmund inequality and a few standard, but rather tedious, moment calculations for Bernoulli processes. We start by exploiting

$$\Pr [|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)}] = \Pr [\langle \mathbf{a}, \mathbf{z} \rangle^2 \geq \theta^2 p(1-p)], \quad (41)$$

because the latter expression is easier to handle. Introducing the nonnegative random variable $S := \langle \mathbf{a}, \mathbf{z} \rangle^2 = \sum_{i,j=1}^n b_i b_j z_i z_j$, we see

$$\mathbb{E} [S] = \sum_{i \neq j} \mathbb{E} [b_i] \mathbb{E} [b_j] z_i z_j + \sum_{i=1}^n \mathbb{E} [b_i^2] z_i^2 = p^2 \langle \mathbf{1}, \mathbf{z} \rangle + p(1-p) \|\mathbf{z}\|_{\ell_2}^2 \geq p(1-p) \quad (42)$$

(recall that each b_i is an independent Bernoulli variable with parameter p). This calculation together with (41) implies

$$\Pr [|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)}] \geq \Pr [S \geq \theta^2 \mathbb{E} [S]]. \quad (43)$$

Since $S \geq 0$ by definition, the requirements for Paley-Zygmund – Theorem 16 – are met and said Theorem implies

$$\Pr [S \geq \theta^2 \mathbb{E} [S]]^2 \geq \frac{(1-\theta^2)^2 \mathbb{E} [S]}{\text{Var}(S) + \mathbb{E} [S]^2}. \quad (44)$$

We have already computed $\mathbb{E} [S]$ in (42), but we still have to compute its variance. We defer this calculation to the very end of this section and for now simply state its result:

$$\text{Var}(S) = 2\mathbb{E} [S]^2 - 2p^4 \langle \mathbf{1}, \mathbf{z} \rangle + 4p^2(1-p)(1-2p) \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i=1}^n z_i^3 + p(1-p)(1-6p(1-p)) \|\mathbf{z}\|_{\ell_4}^4. \quad (45)$$

We now move on to bound these contributions individually by a multiple of $\mathbb{E} [S]^2$. We can omit the second term and obtain

$$\begin{aligned} 4p^2(1-p)(1-2p) \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i=1}^n z_i^3 &\leq 4p^2(1-p)^2 \langle \mathbf{1}, \mathbf{z} \rangle \|\mathbf{z}\|_{\ell_2}^3 = 4p^2(1-p)^2 \langle \mathbf{1}, \mathbf{z} \rangle \leq 4p^2(1-p)^2 \max \{ \langle \mathbf{1}, \mathbf{z} \rangle^2, 1 \} \\ &\leq \frac{2}{p} (p^2 \langle \mathbf{1}, \mathbf{z} \rangle^2 + p(1-p))^2 = \frac{2}{p} \mathbb{E} [S]^2 \end{aligned}$$

for the third term. The fourth term can be bounded via

$$p(1-p)(1-6p(1-p))\|\mathbf{z}\|_{\ell_4}^4 \leq p(1-p)\|\mathbf{z}\|_{\ell_2}^4 \leq \frac{1}{p(1-p)}\mathbb{E}[S]^2.$$

and combining all these bounds implies

$$\text{Var}(S) \leq \left(2 + \frac{2}{p} + \frac{1}{p(1-p)}\right) \mathbb{E}[S]^2 = \frac{3-2p^2}{p(1-p)} \mathbb{E}[S]^2 \leq \frac{3}{p(1-p)} \mathbb{E}[S]^2.$$

Inserting this upper bound into the Paley-Zygmund estimate (44) yields

$$\Pr \left[|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)} \right] \geq \frac{(1-\theta^2)^2 \mathbb{E}[S]^2}{\text{Var}(S) + \mathbb{E}[S]^2} \geq \frac{(1-\theta^2)^2 \mathbb{E}[S]^2}{\left(\frac{3}{p(1-p)} + 1\right) \mathbb{E}[S]^2} \geq \frac{4}{13} p(1-p)(1-\theta^2)^2,$$

as claimed in (25) and (40), respectively. In the last line, we have used $p(1-p) \leq \frac{1}{4}$ for any $p \in [0, 1]$.

Finally, we provide the derivation of Equation (45). We use our knowledge of $\mathbb{E}[S] = p^2 \langle \mathbf{1}, \mathbf{z} \rangle + p(1-p)\|\mathbf{z}\|_{\ell_2}^2$ together with the elementary formula

$$(b_i - p)(b_j - p) = (b_i b_j - p^2) - p b_i - p b_j + 2p^2$$

to rewrite $S - \mathbb{E}[S]$ as

$$\begin{aligned} S - \mathbb{E}[S] &= \sum_{i,j=1}^n b_i b_j z_i z_j - p^2 \sum_{i \neq j} z_i z_j - p \sum_{i=1}^n z_i^2 = \sum_{i \neq j} (b_i b_j - p^2) z_i z_j + \sum_{i=1}^n (b_i^2 - p) z_i^2 \\ &= \sum_{i \neq j} ((b_i - p)(b_j - p) + p b_i + p b_j - 2p^2) z_i z_j + \sum_{i=1}^n (b_i^2 - p) z_i^2 \\ &= \sum_{i \neq j} (b_i - p)(b_j - p) z_i z_j + \sum_{i=1}^n (b_i^2 - p) z_i^2 + p \sum_{i \neq j} b_i z_i z_j + p \sum_{j \neq i} b_j z_j z_i - 2p^2 \sum_{i \neq j} z_i z_j \\ &= \sum_{i \neq j} (b_i - p)(b_j - p) z_i z_j + \sum_{i=1}^n (b_i^2 - p) z_i^2 + 2p \sum_{i,j=1}^n b_i z_i z_j - 2p \sum_{i=1}^n b_i z_i^2 - 2p^2 \sum_{i,j=1}^n z_i z_j + 2p^2 \sum_{i=1}^n z_i^2 \\ &= \sum_{i \neq j} (b_i - p)(b_j - p) z_i z_j + \sum_{i=1}^n (b_i^2 - p) z_i^2 + 2p \sum_{i,j=1}^n (b_i - p) z_i z_j - 2p \sum_{i=1}^n (b_i - p) z_i^2 \\ &= 2 \sum_{i < j} (b_i - p)(b_j - p) z_i z_j + 2p \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i=1}^n (b_i - p) z_i + (1 - 2p) \sum_{i=1}^n (b_i - p) z_i^2. \end{aligned}$$

Here we have exploited symmetry in the first term and $b_i^2 = b_i$ to further simplify that expression. For notational simplicity, it makes sense to define the random variable $\tilde{b}_i := b_i - p$ which obeys $\mathbb{E}[\tilde{b}_i] = 0$ and $\mathbb{E}[\tilde{b}_i^2] = \text{Var}(b_i) = p(1-p)$. Introducing such a notation simplifies the above expression to

$$S - \mathbb{E}[S] = 2 \sum_{i < j} \tilde{b}_i \tilde{b}_j z_i z_j + 2p \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i=1}^n \tilde{b}_i z_i + (1 - 2p) \sum_{i=1}^n \tilde{b}_i z_i^2.$$

Employing the binomial formula $(a + b + c)^2 = a^2 + 2ab + 2ac + b^2 + 2bc + c^2$, we obtain

$$\begin{aligned} \text{Var}(S) = & \mathbb{E} \left[(S - \mathbb{E}[S])^2 \right] = 4 \sum_{i < j} \sum_{k < l} \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \tilde{b}_k \tilde{b}_l \right] z_i z_j z_k z_l + 8p \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i < j} \sum_{k=1}^n \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \tilde{b}_k \right] z_i z_j z_k \\ & + 4(1 - 2p) \sum_{i < j} \sum_{k=1}^n \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \tilde{b}_k \right] z_i z_j z_k^2 + 4p^2 \langle \mathbf{1}, \mathbf{z} \rangle^2 \sum_{i,j=1}^n \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \right] z_i z_j \\ & + 4p(1 - 2p) \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i,j=1}^n \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \right] z_i z_j^2 + (1 - 2p)^2 \sum_{i,j=1}^n \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \right] z_i^2 z_j^2. \end{aligned}$$

Centeredness of \tilde{b} together with the summation constraints ($i < j$) and ($k < l$) implies that summands in the first term vanish, unless $i = k$ and $j = l$. This in turn implies

$$\begin{aligned} 4 \sum_{i < j} \sum_{k < l} \mathbb{E} \left[\tilde{b}_i \tilde{b}_j \tilde{b}_k \tilde{b}_l \right] z_i z_j z_k z_l &= 4 \sum_{i < j} \mathbb{E} \left[\tilde{b}_i^2 \right] \mathbb{E} \left[\tilde{b}_j^2 \right] z_i^2 z_j^2 = 2p^2(1 - p)^2 \sum_{i \neq j} z_i^2 z_j^2 \\ &= 2p^2(1 - p)^2 \left(\sum_{i,j=1}^n z_i^2 z_j^2 - \sum_{i=1}^n z_i^4 \right) = 2p^2(1 - p)^2 (\|\mathbf{z}\|_{\ell_2}^4 - \|\mathbf{z}\|_{\ell_4}^4). \end{aligned}$$

Using a similar argument allows us to conclude that the second and third term must identically vanish (because the index constraints $i < j$ prevents $i = j = k$ and, consequently, at least one index must always remain unpaired). We can exploit $\mathbb{E} \left[\tilde{b}_i \tilde{b}_j \right] = p(1 - p)\delta_{i,j}$ in the remaining terms to conclude

$$\begin{aligned} \text{Var}(S) = & 2p^2(1 - p)^2 (\|\mathbf{z}\|_{\ell_2}^4 - \|\mathbf{z}\|_{\ell_4}^4) + 4p^3(1 - p) \langle \mathbf{1}, \mathbf{z} \rangle^2 \|\mathbf{z}\|_{\ell_2}^2 \\ & + 4p^2(1 - p)(1 - 2p) \langle \mathbf{1}, \mathbf{z} \rangle \sum_{i=1}^n z_i^3 + p(1 - p)(1 - 2p)^2 \|\mathbf{z}\|_{\ell_4}^4. \end{aligned}$$

Slightly rewriting this expression then yields the result presented in (45)

REFERENCES

- [1] Y. Vardi, “Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data,” *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.
- [2] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, “Network Tomography: Recent Developments,” *Statistical Science*, vol. 19, pp. 499–517, 2004.
- [3] J. E. Boyd and J. Meloche, “Evaluation of statistical and multiple-hypothesis tracking for video traffic surveillance,” *Machine Vision and Applications*, vol. 13, no. 5-6, pp. 344–351, 2003.
- [4] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and S. A. S., “Maximum Entropy and the Nearly Black Object,” *Journal of the Royal Statistical Society B*, vol. 54, no. 1, 1992.
- [5] G. Zhang, S. Jiao, X. Xu, and L. Wang, “Compressed sensing and reconstruction with Bernoulli matrices,” in *2010 IEEE International Conference on Information and Automation, ICIA 2010*, 2010, pp. 455–460.
- [6] M. A. Khajehnejad, A. G. Dimakis, W. Xu, and B. Hassibi, “Sparse recovery of nonnegative signals with minimal expansion,” *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 196–208, 2011.
- [7] N. Meinshausen, “Sign-constrained least squares estimation for high-dimensional regression,” *Electronic Journal of Statistics*, vol. 7, no. 1, pp. 1607–1631, 2013.
- [8] M. Slawski and M. Hein, “Sparse recovery by thresholded non-negative least squares,” *Electronic Journal of Statistics*, vol. 7, 2013.
- [9] S. Foucart and D. Koslicki, “Sparse Recovery by Means of Nonnegative Least Squares,” *IEEE Signal Processing Letters*, vol. 21, no. 4, 2014.

- [10] D. L. Donoho and J. Tanner, “Sparse nonnegative solution of underdetermined linear equations by linear programming,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.
- [11] A. M. Bruckstein, M. Elad, and M. Zibulevsky, “On the uniqueness of non-negative sparse & redundant representations,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 796, pp. 5145–5148, 2008.
- [12] S. Foucart and H. Rauhut, “A mathematical introduction to compressive sensing,” *Appl. Numer. Harmon. Anal. Birkhäuser, Boston*, in ..., 2013.
- [13] M. Rudelson and S. Zhou, “Reconstruction from anisotropic random measurements,” *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3434–3447, June 2013.
- [14] R. Kueng and D. Gross, “{RIPless} compressed sensing from anisotropic measurements,” *Linear Algebra and its Applications*, vol. 441, pp. 110 – 123, 2014, special Issue on Sparse Approximate Solution of Linear Systems.
- [15] S. Mendelson, “Learning without concentration,” *J. ACM*, vol. 62, no. 3, pp. 21:1–21:25, Jun. 2015.
- [16] V. Koltchinskii and S. Mendelson, “Bounding the smallest singular value of a random matrix without concentration,” *International Mathematics Research Notices*, vol. 2015, no. 23, pp. 12991–13008, 2015.
- [17] J. A. Tropp, *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*. Cham: Springer International Publishing, 2015, ch. Convex Recovery of a Structured Signal from Independent Random Linear Measurements, pp. 67–101.
- [18] E. J. Candes, T. Strohmer, and V. Voroninski, “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [19] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, “Stable low-rank matrix recovery via null space properties,” *arXiv preprint arXiv:1507.07184*, 2015.
- [20] S. Dirksen, G. Lecué, and H. Rauhut, “On the gap between rip-properties and sparse recovery conditions,” *arXiv preprint arXiv:1504.05073*, 2015.
- [21] Y. Chang, P. Jung, C. Zhou, and S. Stanczak, “Block Compressed Sensing based Distributed Resource Allocation for M2M Communications,” *accepted for International Conference on Acoustics, Speech, and Signal Processing, ICASSP16*, 2016.
- [22] M. Wang, W. Xu, and A. Tang, “A unique “nonnegative” solution to an underdetermined system: From vectors to matrices,” *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1007–1016, 2011.
- [23] R. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, “A Simple Proof of the Restricted Isometry Property for Random Matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, jan 2008.
- [24] E. J. Candes, “The restricted isometry property and its implications for compressed sensing,” *Compte Rendus de l’Academie des Sciences*, vol. 346, no. Paris, Serie I, pp. 589–592, may 2008.
- [25] C. Giraud, S. Huet, and N. Verzelen, “High-Dimensional Regression with Unknown Variance,” *Statistical Science*, vol. 27, pp. 500–518, 2012. [Online]. Available: <http://projecteuclid.org/euclid.ss/1356098553>
- [26] T. Sun and C.-H. Zhang, “Scaled Sparse Linear Regression,” 2011. [Online]. Available: <http://arxiv.org/abs/1104.4595>
- [27] N. Städler, P. Bühlmann, and S. van de Geer, “Rejoinder: l1-penalization for mixture regression models,” *Test*, vol. 19, pp. 209–256, 2010.
- [28] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *SIAM Review*, vol. 57, no. 2, pp. 225–251, 2015.
- [29] R. Kueng, H. Rauhut, and U. Terstiege, “Low rank matrix recovery from rank one measurements,” *Applied and Computational Harmonic Analysis*, pp. –, 2015.
- [30] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Prentice-Hall, 1974.
- [31] E. van den Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.
- [32] J. A. Tropp, “User friendly tools for random matrices. An introduction.” *Preprint*, 2012.
- [33] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2011.